Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Biases incurred from non-random repeat testing of haemoglobin levels in blood donors *Selective testing and its implications*

Ryan K. Chung[*,1,2], **Angela M. Wood** [1,2], and **Michael J. Sweeting** [1,2,3]

[1] Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Worts' Causeway, Cambridge, CB1 8RN, UK

[2] National Institute for Health Research (NIHR) Blood and Transplant Research Unit (BTRU) in Donor Health and Genomics, University of Cambridge

[3] Department of Health Sciences, University of Leicester, Leicester, LE1 7RH, UK

Received zzz, revised zzz, accepted zzz

To help prevent anaemia, it is a requisite for blood donors to undergo a haemoglobin test to ensure levels are not too low before donation. It is therefore important to have an accurate testing device and strategy to ensure donors are not being inappropriately bled. A recent study in blood donors used a selective testing strategy where if a donor's haemoglobin level is below the level required for donation, then another reading is taken and if this occurs again, a third and final reading is used. This strategy can reduce the average number of readings required per donor compared to taking three measurements for all donors. However, the final decision-making measurement will on average be higher than a single measurement. In this paper, a selective testing strategy is compared against other strategies. Individual-level biases are derived for the selective strategy and are shown to depend on how close a donor's true haemoglobin level is to the donation threshold and the magnitude of error in the testing device. A simulation study was conducted using the distribution of haemoglobin levels from a large donor population to investigate the effects different strategies have on population performance. We consider scenarios based on varying the measurement device bias and error, including differential biases that depend on the underlying haemoglobin level. Discriminatory performance is shown to be affected when using the selective testing strategies, especially when measurement error is large and when differential bias is present in the device. We recommend that the average of a number of readings should be used in preference to selective testing strategies if multiple measurements are available.

Key words: Blood donation; Diagnostic performance; Multiple measurements; Selective testing; Simulation study;

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX

1 Introduction

Blood donations are an integral part of many health services across the world (`http://www.who.int/mediacentre/factsheets/fs279/en`). To help prevent anaemia, it is a requisite for blood donors to undergo a haemoglobin (Hb) test to ensure levels are not too low before donation. In most European countries the Hb threshold levels for donation are set at 135g/L for males and 125g/L for females. A number of point-of-care screening tests are available, including measuring Hb in capillary or venous blood samples, for example with a HemoCue® device (`http://www.hemocue.com/en/solutions/hematology`) or using a non-invasive device that uses technology such as spectroscopy to measure Hb levels without taking a blood sample. Whilst the measuring performance of a device is important, the strategy employed when using these devices is

*Corresponding author: e-mail: rkyc2@medschl.cam.uk, Phone: +44-(0)1223-747217, Fax: +44-(0)1223-748658

arguably as important. Singh et al. (2015) and Hiscock et al. (2014) suggest using an average of multiple Hb measurements to increase precision in the estimated Hb level and therefore increase accuracy in the decision of whether the donor should donate or not.

Recently in a paper by Baart et al. (2016), a strategy was used such that if a donor is tested and their measured Hb level is below a sex-specific donation threshold required for donation, then a second measurement should be taken and if that is below the threshold then a third measurement should be taken. The final measurement taken becomes the decision-making measurement for the donor and is used to determine whether they should donate. Such a strategy can reduce the average number of measurements taken per donor compared with measuring all donors three times (and using the average to make a decision), and can result in fewer deferrals for donors who are close to the donation threshold. The number of measurements taken depend on the donor's measured Hb level. For example a low-risk donor with high Hb may need to be measured only once to ensure that they can donate above the threshold, whilst a donor close to the donation threshold might be measured at most three times. Therefore in a general population, fewer measurements are taken compared with taking three measurements for all donors and the average time needed to measure a donor's Hb would be reduced.

However, assessing the performance of the testing device is compromised by the use of the testing strategy. The selective testing strategy described is akin to what previous papers have described as "sequential testing", which could lead to sequential testing bias. For example, in clinical trials, decisions to continue or not are made at the end of each trial phase and can lead to sequential biases that affect evidence syntheses as shown by Kulinskaya et al. (2016) and Denne (2000). Similar problems arise in group sequential clinical trials where an interim decision to stop or extend one or more trial arm can bias the final treatment effect estimate, as shown by Whitehead (1986), Liu et al. (2004), Koopmeiners et al. (2012) and Choodari-Oskooei et al. (2013). The selective testing strategy described for blood donors presents an interesting extension of a sequential bias problem where decisions to proceed (take a further measurement) or not are made for each individual donor.

A selective testing strategy in blood donor studies may have implications on the individual-level bias which may also have implications exhibited at the population level, such as an increase in the number of donors who are bled below the Hb threshold. The reported sensitivity and specificity of a device may also be compromised if such a testing strategy is implemented, and may partly be the reason why a recent study of a number of devices estimated very low sensitivities ranging from 3.5% to 37% (Baart et al., 2016). Studies by Pagliaro et al. (2014), Clippel et al. (2017) and Ahankari et al. (2016) did not highlight the use of this strategy and showed higher sensitivity levels ranging from 13% to 63%. To investigate this further we quantify the bias of a selective testing strategy as a function of a donor's true Hb level and the measurement error in the device at the individual-level, and explore in a simulation study the effect the selective testing strategy has on the population-level by measuring statistics such as the bias in estimating the mean Hb, and the sensitivity and specificity of the decision-making measurement. These are of particular interest to a blood donation service as the bias can inform whether the device is under or overestimating relative to a "gold-standard", whilst the sensitivity and specificity can determine the proportion of donors correctly deferred or bled.

In this paper we compare the selective testing strategy to a strategy that uses an average of a pre-defined number of measurements and a strategy that uses selective testing based on the average of previous measurements to define the decision-making measurement. The testing strategies are presented formally in Section 2. In Section 3, the bias of the selective strategies for a given range of true Hb levels close to the Hb donation threshold is quantified. Section 4 describes a simulation study for a population of donors where the overall bias in the reported mean Hb level, coverage of the confidence intervals, variance, and the sensitivity and specificity of the testing strategies in identifying donors below and above the donation

threshold are estimated. Finally in Section 5, recommendations are made regarding the different strategies for repeat testing of Hb levels.

2 Testing Strategies

The first strategy uses selective testing (ST) to obtain a final decision-making measurement for each donor. Given the results of an initial Hb measurement, a second measurement is taken if and only if the initial measurement is below a sex-specific donation threshold. A third measurement is taken if and only if the second measurement is below the sex-specific donation threshold and this continues up to a maximum of K measurements. The final measurement that is taken is used to define the donor's decision-making measurement, which is used for determining whether the individual is bled or deferred. When $K = 1$ this strategy reduces to a single measurement test. The second strategy uses the average of K measurements for all donors (labelled the Average Measurements (AM) strategy), where K is pre-specified. The third strategy is similar to the selective testing strategy, which takes a new measurement if and only if the previous measurement is below the threshold, but an average of the measurements taken is used to define the donor's final decision-making measurement, and is therefore a combination of the ST and AM strategies (labelled the STAM strategy). The STAM strategy was investigated because, like the ST strategy, it benefits from needing on average fewer measurements per donor than the AM strategy but may also be less biased than the ST strategy.

Let $A_i^{(K)}$, $B_i^{(K)}$ and $C_i^{(K)}$ represent the decision-making Hb level from the ST, AM and STAM testing strategies, respectively, for the i^{th} donor using a maximum of K measurements and let y_{ik} be the k^{th} $(k = 1, ..., K)$ measurement for the i^{th} $(i = 1, ..., n)$ donor. Let T denote the Hb threshold for donation (e.g. 135g/L in males). Then the three key testing strategies are formally defined as follows:
(ST) Selective testing strategy with a maximum of K measurements:

$$A_i^{(K)} \sim \begin{cases} y_{i1} & \text{if } y_{i1} \geq T \\ y_{i2} & \text{if } y_{i1} < T \text{ and } y_{i2} \geq T \\ & \vdots \\ y_{iK-1} & \text{if } y_{i1} < T, \ldots, y_{iK-2} < T \text{ and } y_{iK-1} \geq T \\ y_{iK} & \text{if } y_{i1} < T, \ldots, y_{iK-1} < T \end{cases}$$

(AM) Average of K measurements:

$$B_i^{(K)} = \frac{1}{K}(y_{i1} + \ldots + y_{iK})$$

(STAM) Selective testing and averaging strategy with a maximum of K measurements:

$$C_i^{(K)} \sim \begin{cases} y_{i1} & \text{if } y_{i1} \geq T \\ \frac{1}{2}(y_{i1} + y_{i2}) & \text{if } y_{i1} < T \text{ and } y_{i2} \geq T \\ & \vdots \\ \frac{1}{K-1}(y_{i1} + \ldots + y_{iK-1}) & \text{if } y_{i1} < T, \ldots, y_{iK-2} < T \text{ and } y_{iK-1} \geq T \\ \frac{1}{K}(y_{i1} + \ldots + y_{iK}) & \text{if } y_{i1} < T, \ldots, y_{iK-1} < T \end{cases}$$

Seven variations of the above testing strategies are discussed in this paper; a selective testing strategy with a maximum of two or three measurements, an averaging strategy with one, two or three measurements, and

the selective testing and averaging strategy with a maximum of two or three measurements. These will be denoted as ST2, ST3, AM1, AM2, AM3, STAM2 and STAM3.

3 Bias of testing strategies under Normally distributed measurement error

3.1 Derivation of individual-level bias

In this section the individual-level bias of the decision-making Hb level for each testing strategy under a Normal measurement error model is quantified. Suppose that a point-of-care Hb test device is unbiased so that, given a true Hb level μ_i from individual i, observed measurements can be described as realisations from a Normal distribution with mean μ_i and standard deviation σ_e. Since repeat measurements are taken together over a short space of time we shall assume that the variation in Hb levels within an individual is due to measurement error from the testing device and not from changes in the underlying phenotype, as might be the case for example with blood pressure. For an individual i, let each measurement, conditional on μ_i, be independent and identically distributed with

$$Y_{ik} \sim N(\mu_i, \sigma_e^2)\ ; k = 1, ..., K.$$

Therefore for all testing strategies the Y_{ik} are assumed independent from one another.

For the AM testing strategy with K measurements it is trivial to see that the estimated Hb level is unbiased:

$$\mathrm{E}\left(B_i^{(K)}\right) = \mathrm{E}\left(\frac{\sum_{k=1}^{K} Y_{ik}}{K}\right) = \mu_i$$

For the ST testing strategy, and dropping the subscript i from this point onwards, $\mathrm{E}\left(A^{(K)}\right)$ and the bias $\mathrm{E}\left(A^{(K)}\right) - \mu$ can be obtained as follows. First focussing on the case where the maximum number of measurements is two (ST2) to obtain $\mathrm{E}\left(A^{(2)}\right)$:

$$\mathrm{E}(A^{(2)}) = \mathrm{E}(Y_1|Y_1 \geq T)p(Y_1 \geq T) + \mathrm{E}(Y_2|Y_1 < T)p(Y_1 < T)$$

Note that the first term of the equation is a truncated normal (for y_1) whilst the second part is a conditional normal (of y_2 given Y_1). To make calculations simpler further on, each measurement Y_k is standardised such that $Z_k = \frac{Y_k - \mu}{\sigma_e} \sim N(0, 1)\ ; k = 1, ..., K$.

Then the expectation of $A^{(2)}$ can be written as

$$\mathrm{E}(A^{(2)}) = [\sigma_e \mathrm{E}(Z_1|Z_1 \geq T_z) + \mu]p(Z_1 \geq T_z) + [\sigma_e \mathrm{E}(Z_2|Z_1 < T_z) + \mu]p(Z_1 < T_z)$$

where $T_z = \frac{T - \mu}{\sigma_e}$

Using results of the truncated and conditional normal distributions as shown by Maddala (1983), $\mathrm{E}(Z_1|Z_1 > a) = \frac{\phi(a)}{1-\Phi(a)}$ and $\mathrm{E}(Z_1|Z_2 < a) = -\rho\frac{\phi(a)}{\Phi(a)}$, where Z_1 and Z_2 are both standard normal, $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard Normal density and distribution functions, respectively, and ρ is the correlation between Z_1 and Z_2. Since each y_k, and therefore each z_k, are independent from one another, conditional on μ, and the correlation $\rho = 0$, the expectation for $A^{(2)}$ is therefore:

$$\mathrm{E}(A^{(2)}) = \left(\sigma_e \frac{\phi(T_z)}{1 - \Phi(T_z)} + \mu\right)(1 - \Phi(T_z)) + \mu\Phi(T_z) = \mu + \sigma_e\phi(T_z)$$

Working through the strategy ST3 in a similar manner, the expectation is:

$$\begin{aligned}\mathrm{E}\left(A^{(3)}\right) = &\ \mathrm{E}(Y_1|Y_1 \geq T)\, p(Y_1 \geq T) \\ &+ \mathrm{E}(Y_2|Y_1 < T, Y_2 \geq T)\, p(Y_1 < T, Y_2 \geq T) \\ &+ \mathrm{E}(Y_3|Y_1 < T, Y_2 < T)\, p(Y_1 < T, Y_2 < T)\end{aligned}$$

which due to independence simplifies to:

$$\begin{aligned}\mathrm{E}\left(A^{(3)}\right) = &\ \mathrm{E}(Y_1|Y_1 \geq T)\, p(Y_1 \geq T) \\ &+ \mathrm{E}(Y_2|Y_2 \geq T)\, p(Y_1 < T) p(Y_2 \geq T) \\ &+ \mathrm{E}(Y_3)\, p(Y_1 < T) p(Y_2 < T)\end{aligned}$$

Standardising and using results of the truncated and conditional normal distributions, as before, results in the following equation for the expectation of the ST3 strategy:

$$\mathrm{E}\left(A^{(3)}\right) = \mu + \sigma_e \phi(T_z)(1 + \Phi(T_z)) \geq \mu$$

The bias in the Hb level estimated from the ST strategy is given by $\mathrm{E}\left(A^{(K)}\right) - \mu$. The individual-level bias for ST2 and ST3 is therefore:

$$\begin{aligned}\text{Bias}\left(A^{(2)}\right) &= \sigma_e \phi(T_z) \geq 0 \\ \text{Bias}\left(A^{(3)}\right) &= \sigma_e \phi(T_z)(1 + \Phi(T_z)) \geq 0\end{aligned}$$

Generalising the amount of bias for any selective testing strategy where the maximum number of measurements is K gives the following result:

$$\text{Bias}\left(A^{(K)}\right) = \sigma_e \phi(T_z) \left(\sum_{k=0}^{K-2} \Phi(T_z)^k\right) \text{ for } K \geq 2$$

In a similar manner, generalising the amount of bias for any selective testing and averaging (STAM) strategy with a maximum of K measurements gives the following result:

$$\text{Bias}\left(C^{(K)}\right) = \sigma_e \phi(T_z) \left(\sum_{k=0}^{K-2} \frac{1}{k+2} \Phi(T_z)^k\right) \text{ for } K \geq 2$$

3.2 Calculation of individual-level bias for different μ and σ_e

To understand the effects of how the individual-level bias for the selective testing strategies ST2, ST3, STAM2 and STAM3 change with varying levels of measurement error, σ_e, and underlying Hb level μ, the bias was plotted against the difference in μ from the Hb threshold. Figure 1 shows the bias at the individual-level as the measurement error is increased from 1g/L, to 6g/L, to 11g/L.

The maximum amount of bias exhibited for testing strategies ST2 and STAM2 is equivalent to $\sigma_e\,(\phi(0))$ and $\sigma_e\left(\frac{1}{2}\phi(0)\right)$ respectively, approximately $0.40\sigma_e$ and $0.20\sigma_e$, and occurs when the true Hb level is at

the threshold. In contrast, the maximum bias for ST3 and STAM3 is approximately $0.62\sigma_e$ and $0.27\sigma_e$ respectively. To determine the location of the maximum bias, a numerical approximation was calculated by plotting the bias for multiple values of σ_e and using its maximum to determine the difference from the Hb threshold for when this occurs. For ST3 and STAM3 respectively, these were $-0.25\sigma_e$ g/L and $-0.19\sigma_e$ g/L from the threshold.

4 Simulation Study

4.1 Data Generation

A simulation study was conducted in `R` to assess the bias and coverage in estimating the population mean Hb level, and the diagnostic performance of the testing strategies on a population with an Hb distribution typical of presenting blood donors. This was done to understand what effect the individual-level biases have on blood donations in a general population. The data for the simulation study was generated by performing 1000 simulations, with 2500 donors in each simulation. The number of donors was chosen to closely represent the number of donors analysed in previously published medium-sized Hb-testing studies such as those by Baart et al. (2016), Pagliaro et al. (2014) and Ziemann et al. (2011).

For each donor, their true Hb level μ_i was drawn from a Normal distribution that aimed to approximate a distribution of Hb values measured using a "gold standard" haematology analyser in 13,413 males and 15,081 females presenting donors in the COMPARE study (`www.comparestudy.org.uk`). The empirical mean and standard deviation from the COMPARE study was 151g/L and 10g/L for males and 135g/L and 10g/L for females, respectively, whilst the prevalence of low Hb in the population was 5.2% for males and 13.3% for females (using the standard thresholds of $<$135g/L for males and $<$125g/L for females). A simulated data set using the empirical mean and standard deviation was generated. Since the distributions of Hb seen in the COMPARE population of donors are fairly similar in shape for either sex, results in the simulation study were shown for males only. An approximation of Hb values was used due to the confidentiality of data.

Three point-of-care Hb measurements were independently drawn for each donor from a Normal distribution with mean $\mu_i + b_i$ and measurement error σ_e, where b_i represents bias in the measurement device for the i^{th} donor. Measurements based on each of the seven strategies described in Section 2 were obtained and the difference between the measurement based on each strategy and the donor's true Hb level was calculated. The bias of the mean estimated Hb level in the population under each testing strategy was calculated for each simulation and the mean bias and coverage of the 95% confidence intervals for the mean Hb level were summarised over the simulations.

The sensitivity, defined as the proportion of donors with a "gold-standard" measurement below the threshold who also had a decision-making measurement below the threshold, and the specificity, defined as the proportion of donors with a "gold-standard" measurement above the threshold who also had a decision-making measurement above the threshold, were calculated for each testing strategy. The confidence intervals for sensitivity and specificity were calculated using a binomial exact test. The area under the receiver operating characteristic curve (AUC) was used to quantify the discriminatory ability for each strategy. An empirical curve was generated by varying the Hb threshold in increments of 1g/L, whilst the AUC was calculated using the trapezoidal rule (Robin et al., 2011). Performance metrics were then averaged across the 1000 simulations.

4.2 Scenarios

The simulation study was conducted for different values of the measurement error σ_e. A plausible range was chosen using data on 3704 male donors in COMPARE who had three repeated measurements of Hb at the same visit. The mean measurement error in this population was 5.5g/L (standard deviation 4.6g/L). Hence the measurement error in the simulation study varied from 1g/L up to 11g/L. Each of the repeated measurements were automatically taken seconds apart from one another and so within-person variability of Hb levels can be assumed to be negligible. The main scenario considered unbiased measurement devices (e.g $b_i = 0 \, \forall i$). Three further scenarios that introduced bias in the measurement device in different forms were conducted.

The first scenario considered systematic bias in the measurement device such that $b_i = c$ for all i, where c was allowed to vary between -2g/L and +2g/L. Hence the device's bias is constant for all individuals and independent from an individual's Hb measurement. This aimed to show how the sensitivity and specificity of a point-of-care device can be affected by systematic bias. The second scenario varied the amount of bias in the testing device between individuals. This was done to mimic what might be experienced in a real world scenario, such as temperature changes that might affect day-to-day, and hence donor-to-donor, variation. The device bias for each donor was simulated from a Normal distribution such that $b_i \sim N(0, \sigma_b^2)$, where σ_b was varied between 0g/L and 5g/L. Note $\sigma_b = 0$ is equivalent to the main scenario where the device was unbiased for all individuals. The third scenario allowed bias in the device to be defined by a function of an individual's true Hb level to impose differential bias. The gradient of the differential bias (the increase in bias per 1g/L increase in true Hb) was then varied. The bias for the i^{th} donor is given by $b_i = f(\mu_i) = \alpha + \beta\mu_i$, where α is the intercept and β is the gradient of the differential bias. The functions were centred around the mean of the population's true Hb level which keeps the mean device bias equal to zero. β is varied between -0.5 and 0.5 and when the gradient is equal to zero, the bias is equivalent to an unbiased device, i.e. the main scenario. The latter two scenarios were conducted since a point-of-care device's AUC is not sensitive to systematic bias but is affected by differential bias (Pfeiffer, 2017).

A further scenario altered the prevalence of low Hb by changing the baseline Hb distribution. This was done by shifting each individual's Hb level, and thus the mean Hb, by a fixed amount between -2g/L and 2g/L. By keeping the same Hb threshold the prevalence of low Hb is altered. The scenario was investigated due to different populations having different levels of prevalence of low Hb.

5 Results

Population-level results for the simulation study main scenario based on an unbiased measurement device are first shown before considering biases in the measurement device.

5.1 An unbiased measurement device

As the measurement error of the point-of-care device increases, the sensitivity and specificity drops amongst all the testing strategies (Figure 2). The selective testing strategies (ST2 and ST3) are particularly affected and give far worse sensitivity for all levels of measurement error, although due to the positive bias exhibited in these strategies, the specificity is high. Strategies that take an average of two or three measurements give better performance than the single measurement strategy for both sensitivity and specificity. The STAM strategies have a sensitivity that lies between the AM and ST strategies.

The AUC is relatively high for all strategies when the standard deviation was $\leq$7g/L (Table 1). However, as the standard deviation increases, the selective testing strategy with a maximum of three measurements

gives the lowest AUC. With a standard deviation of 11g/L, there is a 10% relative difference in AUC between the testing strategies ST3 and AM3. As expected, the STAM strategies give discriminative performance that lies between the AM and ST strategies.

The mean bias in the estimated population Hb level in males is as high as 2.3g/L under strategy ST3 when the measurement error reaches 11g/L (Figure S1 and Table 1). These biases are not as extreme as those estimated for an individual in Section 3 as the majority of donors in the population have Hb levels sufficiently far away from the sex-specific donation threshold, which reduces the possibility of being misclassified as having low or sufficient Hb. For example, 75% of male donors have Hb levels $\geq$10g/L away (in absolute terms) from the donation threshold. The average measurement (AM) testing strategies are shown to be unbiased for all levels of measurement error.

The coverage of the 95% confidence interval for the mean Hb level in the population is shown in Table 1, and also Figure S2 by the amount of measurement error present in the point-of-care test. The AM1, AM2 and AM3 testing strategies have coverage equal to the nominal coverage rate (e.g. 95 per cent) irrespective of measurement error, whilst the selective testing strategies (ST2, ST3, STAM2 and STAM3) exhibit under-coverage and hence the confidence intervals will contain the true population mean less than 95% of the time. The coverage of ST3 is below 50% when the measurement error is $>$5g/L.

The variance of each of the testing strategies is shown in Figure S3. When the measurement error increases it is trivial to see that the AM testing strategies has a variance equal to $\frac{\sigma_e^2}{K}$ for K measurements. The variance for the ST and STAM testing strategies decreases marginally when more measurements are taken, however they are all similar regardless of how many measurements are taken. The reason for this is that the majority of donors would only have one measurement if a ST or STAM strategy was implemented, and thus the variances are all similar. Comparing the ST and STAM testing strategies with the AM testing strategies show that the AM2 and AM3 testing strategies exhibit smaller variance for the full range of measurement error shown, with the AM1 strategy showing a larger variance as it only takes one measurement for all individuals.

Table 1 also shows that the mean standard error and the empirical standard error are similar across the simulation scenarios and for all strategies suggesting that the drop in coverage for the selective testing strategies arises due to the bias in the estimated population mean. The average number of measurements taken per donor are equivalent for the ST and STAM strategies. Compared to AM3 where three measurements are taken for all donors, both the ST3 and STAM3 strategies reduce the average number of measurements per donor by 1.8.

5.2 A systematically biased measurement device

A plot of the sensitivity, specificity and AUC for each strategy is shown in Figure 3, when systematic bias is introduced into the point-of-care device. The figure shows that a device that generally overestimates Hb can cause a severe decrease in the sensitivity whilst marginally increasing the specificity. For example given a low measurement error of 1g/L, a 2g/L increase in bias reduces the sensitivity from 90% to 60% for ST2 whilst increasing the specificity by 2%. Likewise if a device underestimates Hb, the opposite is true such that it increases the sensitivity whilst reducing the specificity. The changes in sensitivity and specificity are less pronounced when using the AM testing strategy. The AUC is generally unaffected by systematic bias, however there is a small decrease in AUC performance for both the ST and STAM testing strategies when systematic bias is positive.

5.3 Differential bias in measurement device

A plot of the sensitivity, specificity and AUC for each strategy when varying the between-person standard deviation in the point-of-care device bias is shown in Figure 4. The figure shows that increasing the between-person variability in bias reduces the AUC performance for any given strategy when the measurement error is low. However, as the measurement error becomes larger and starts to dominate, between-person variability has less of an effect. This can also be seen in the sensitivity as it converges the larger the measurement error, and to a lesser effect for the specificity.

Figure 5 shows a plot of the three differential biases imposed based on true Hb level and how these affect the sensitivity, specificity and AUC for each strategy. When the gradient of the bias is negative (i.e. the point-of-care device is overestimating Hb for lower true Hb levels whilst underestimating for larger true Hb levels) the sensitivity decreases relative to when the gradient is 0. Conversely the specificity is higher when the gradient is negative. A sensitivity of 0% is seen when the measurement error is low and this is due to the differential bias overestimating all individuals such that no individual is correctly identified as a low Hb individual. The converse is true when the gradient of the bias is positive (i.e. the point-of-care device is overestimating Hb for higher true Hb levels whilst underestimating for lower true Hb levels); the sensitivity increases whilst the specificity decreases relative to when the gradient is 0. Therefore when the gradient is positive the AUC increases relatively whilst when the gradient is negative the AUC decreases dramatically.

5.4 Changing the baseline Hb distribution

The mean bias, sensitivity and specificity exhibited for the testing strategies is shown in Figure S4 for different levels of true low Hb prevalence. Mean bias increases with prevalence and there is a greater absolute difference in mean bias when the measurement error is large, for all selective testing strategies. The figure shows that a change in prevalence does not drastically affect the magnitude of the mean bias, with the greatest differences in mean bias occurring when the measurement error is very large. For example, when the measurement error is at 11g/L the mean bias for the testing strategy ST3 is 2.3g/L at the baseline prevalence of 5.3%. This increases by 0.4g/L when the prevalence increases to 8.1%, and decreases by 0.4g/L when the prevalence decreases to 3.4%. The STAM testing strategies are less affected by prevalence whilst the AM testing strategies all remain unbiased regardless of the prevalence rate. The sensitivity and specificity is also relatively robust to changes in the prevalence for all strategies and amounts of measurement error.

6 Discussion

This paper has shown that a selective testing (ST) strategy and a selective testing and averaging (STAM) strategy can overestimate the decision-making Hb level at the individual-level, especially for individuals close to the donation threshold. This has a consequence at the population-level where more individuals are incorrectly bled below the threshold. The overestimation of the decision-making Hb level can adversely affect the proportion of donors correctly identified as having a low Hb level (sensitivity) whilst marginally increasing the proportion of donors correctly identified as having a sufficient Hb level to donate (specificity).

Discriminatory performance is affected when using the selective testing strategies, whilst taking an average of multiple measurements increases precision and thus increases discriminatory performance. Further analyses, which changed the population prevalence, show that the prevalence of low Hb in the population can affect the population-level mean bias by a small amount, however sensitivity and specificity are largely unaffected by the changes in prevalence. Likewise changing the systematic bias of the device

shows that the sensitivity and specificity can be affected. For example, using the AM1 and assuming a measurement error of 11g/L, the sensitivity went from 68% to 60% when the bias of the point-of-care device went from 0g/L to 2g/L.

The simulations show that a decrease in discriminative performance (as assessed using the AUC) is dramatic when there is differential bias in the point-of-care device used, especially when the point-of-care device overestimates low Hb and underestimates high Hb. This loss of discriminative performance is further exacerbated by the testing strategy used.

Baart et al. (2016) recently used a selective testing strategy with a maximum of three measurements (ST3) to assess the performance of three point-of-care devices in a population of blood donors in the Netherlands. The sensitivity of these devices was found to be very low. The authors comment that results were comparable with a single measurement strategy (AM1). This is contradictory to the findings in this paper, which suggests sensitivity can suffer as a result of using a selective testing strategy.

It is recommended that a selective testing (ST) strategy is not used in practice due to the biases incurred for donors close to the Hb threshold, and that an average of multiple measurements is the best strategy to use. However, a STAM2 strategy may be considered as an alternative to the AM1 strategy, as it can increase specificity and offer similar AUC performance but at a cost of some bias close to the threshold, a decrease in sensitivity and overall coverage.

Acknowledgements We gratefully acknowledge the participation of all COMPARE volunteers. We thank the COMPARE study co-ordination teams at the University of Cambridge and at NHS Blood and Transplant (NHSBT), including the blood donation staff at the 10 mobile centres, for their help with COMPARE participant recruitment and study fieldwork. The COMPARE academic coordinating centre receives core support from the UK Medical Research Council (MR/L003120/1), British Heart Foundation (RG/13/13/30194) and the UK National Institute for Health Research Cambridge Biomedical Research Centre. The COMPARE study is funded by NHSBT and the NIHR Cambridge Biomedical Research Centre and has been supported by the NIHR-BTRU in Donor Health and Genomics (NIHR BTRU-2014-10024) at the University of Cambridge in partnership with NHSBT.

Conflict of Interest

The authors have declared no conflict of interest.

References

Ahankari, A. S., Fogarty, A. W., Tata, L. J., Dixit, J., and Myles, P. R. (2016). Assessment of a non-invasive haemoglobin sensor nbm 200 among pregnant women in rural India. *BMJ Innovations*, 2:70–77.

Baart, A. M., de Kort, W. L., van den Hurk, K., and Pasker-de Jong, P. (2016). Hemoglobin assessment: precision and practicability evaluated in the Netherlands—the HAPPEN study. *Transfusion*, 56:1984–1993.

Choodari-Oskooei, B., Parmar, M. K., Royston, P., and Bowden, J. (2013). Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcome. *Trials*, 14(1):23.

Clippel, D., Heddegem, L., Vandewalle, G., Vandekerckhove, P., and Compernolle, V. (2017). Hemoglobin screening in blood donors: a prospective study assessing the value of an invasive and a noninvasive point-of-care device for donor safety. *Transfusion*, 57:938–945.

Denne, J. S. (2000). Estimation following extension of a study on the basis of conditional power. *Journal of Biopharmaceutical Statistics*, 10:131–144.

Hiscock, R., Simmons, S., Carstensen, B., and Gurrin, L. (2014). Comparison of massimo pronto-7 and hemocue Hb 201+ with laboratory haemoglobin estimation: a clinical study. *Anaesthesia and Intensive Care*, 42:608.

Koopmeiners, J. S., Feng, Z., and Pepe, M. S. (2012). Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Statistics In Medicine*, 31(5):420–435.

Kulinskaya, E., Huggins, R., and Dogo, S. H. (2016). Sequential biases in accumulating evidence. *Research Synthesis Methods*, 7:294–305.

Liu, A., Troendle, J. F., Yu, K. F., and Yuan, V. W. (2004). Conditional maximum likelihood estimation following a group sequential test. *Biometrical Journal*, 46(6):760–768.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge University Press.

Pagliaro, P., Belardinelli, A., Boko, V., Salamon, P., Manfroi, S., and Tazzari, P. L. (2014). A non-invasive strategy for haemoglobin screening of blood donors. *Blood Transfusion*, 12:458.

Pfeiffer, R. M. and Gail, M. H. (2017). *Absolute Risk: Methods and Applications in Clinical Management and Public Health*. CRC Press.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(1):77.

Singh, A., Dubey, A., Sonker, A., and Chaudhary, R. (2015). Evaluation of various methods of point-of-care testing of haemoglobin concentration in blood donors. *Blood Transfusion*, 13:233.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73(3):573–581.

Ziemann, M., Lizardo, B., Geusendam, G., and Schlenke, P. (2011). Reliability of capillary hemoglobin screening under routine conditions. *Transfusion*, 51:2714–2719.

7 Figures and Tables

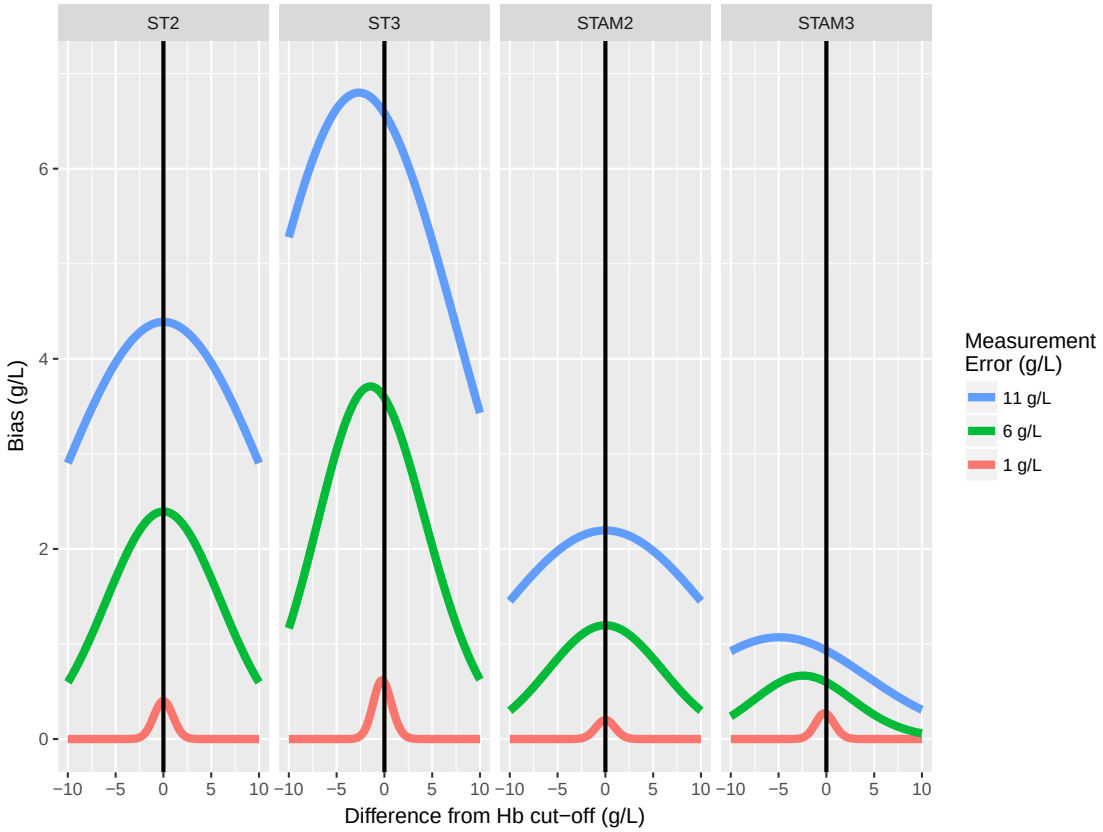

Figure 1 Bias exhibited at the individual-level for the selective testing strategies ST2, ST3, STAM2 and STAM3 on a range of true Hb levels ± 10 g/L from the threshold T according to the amount of measurement error; 1g/L in red, 6g/L in green and 11g/L in blue.

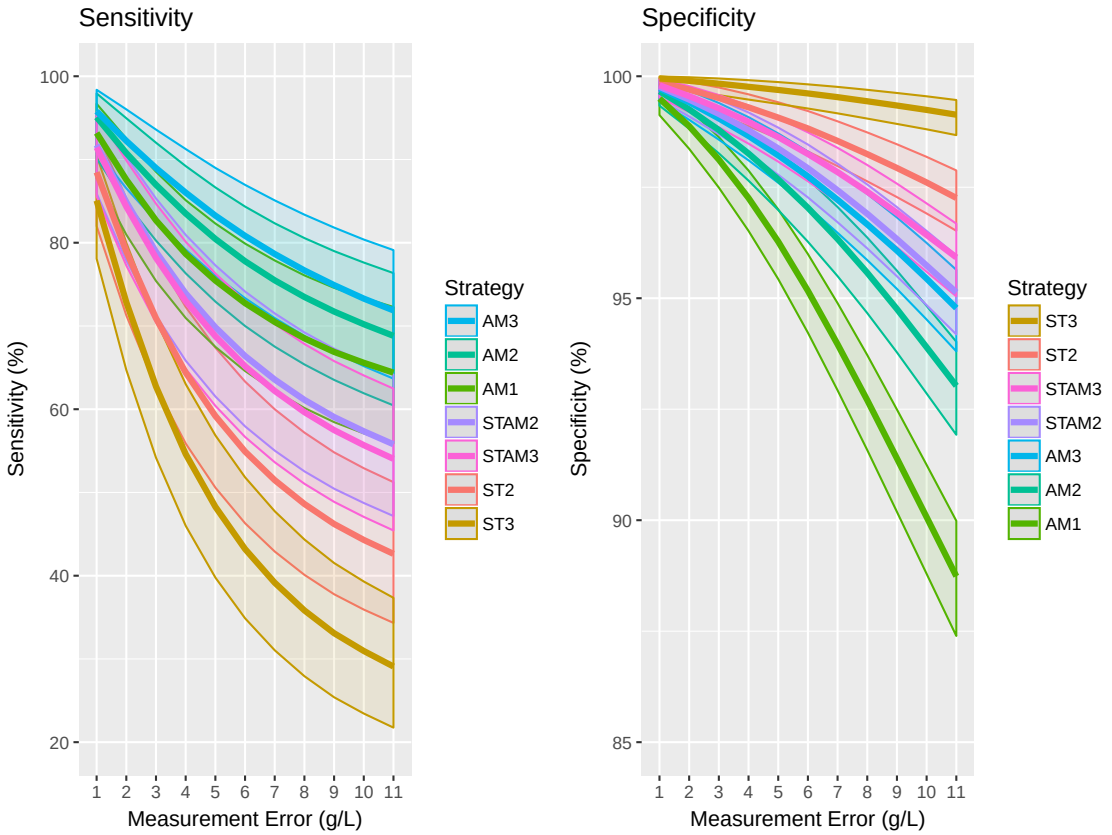

Figure 2 Sensitivity and specificity with 95% binomial exact CI bands for each testing strategy according to the amount of measurement error.

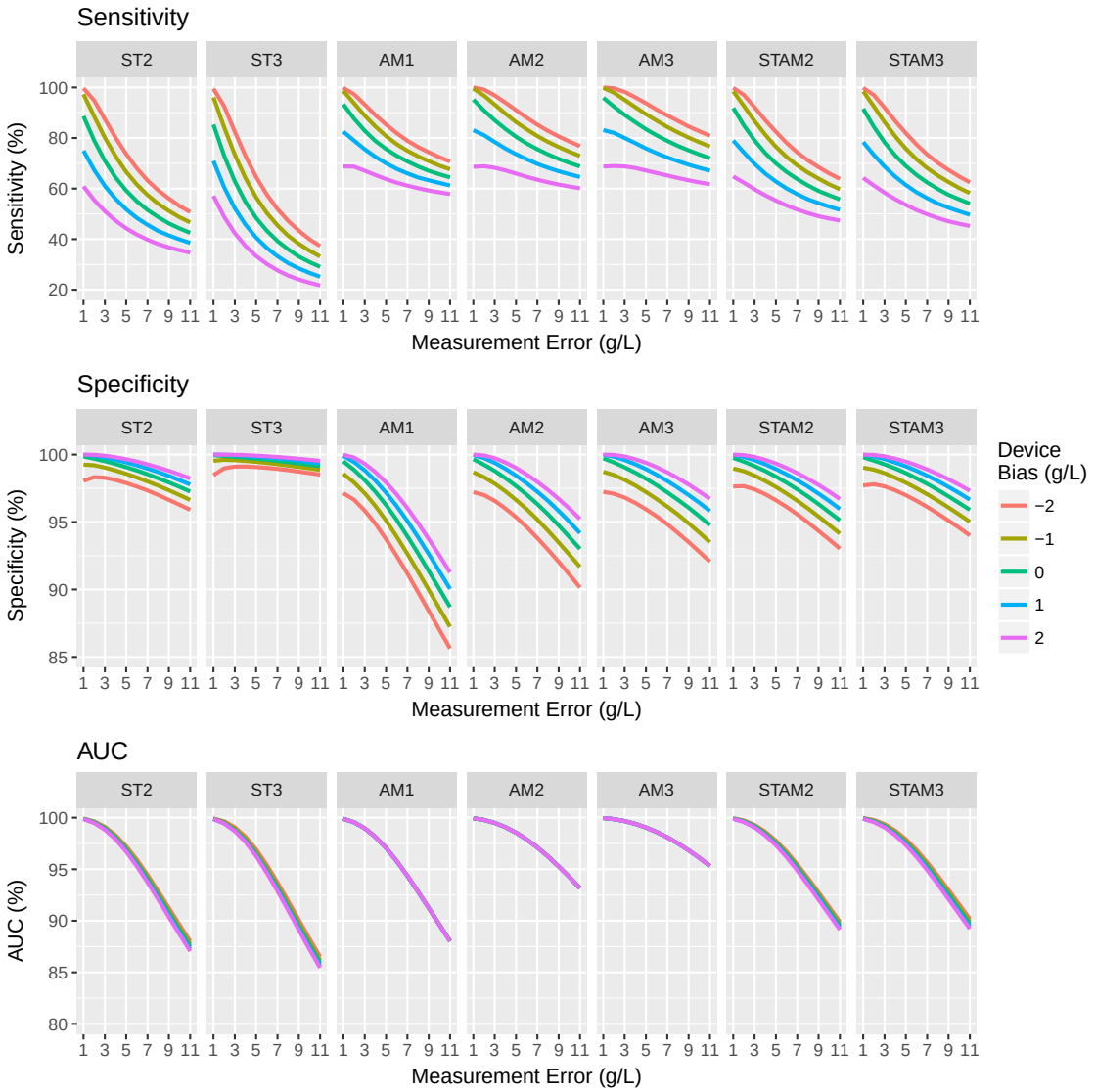

Figure 3 Sensitivity, specificity and AUC exhibited for each testing strategy according to the amount of systematic bias in the measurement device and the measurement error.

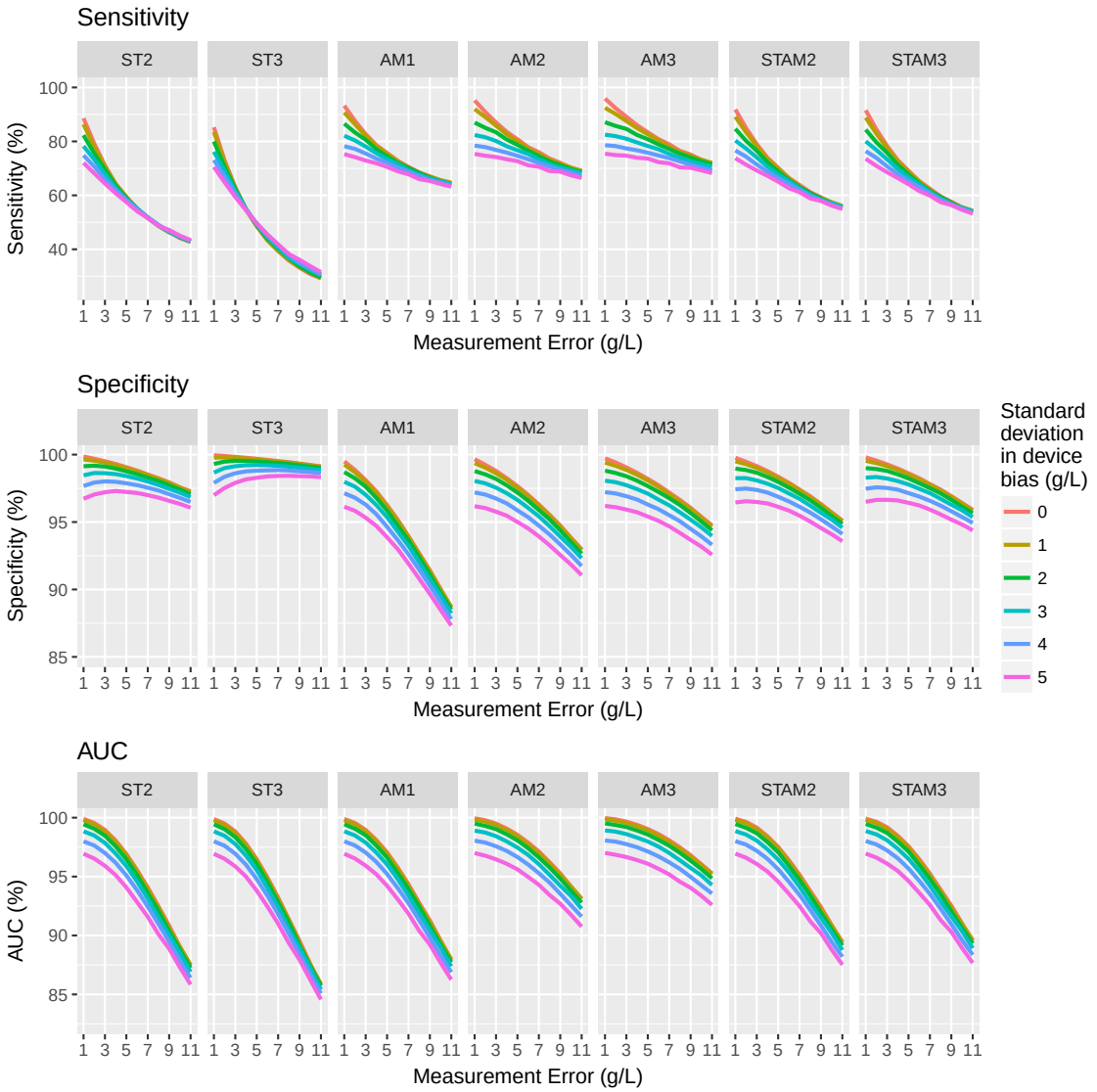

Figure 4 Sensitivity, specificity and AUC exhibited for each testing strategy according to the amount of between-person variability in the measurement device bias, given the average device bias is 0, and the measurement error.

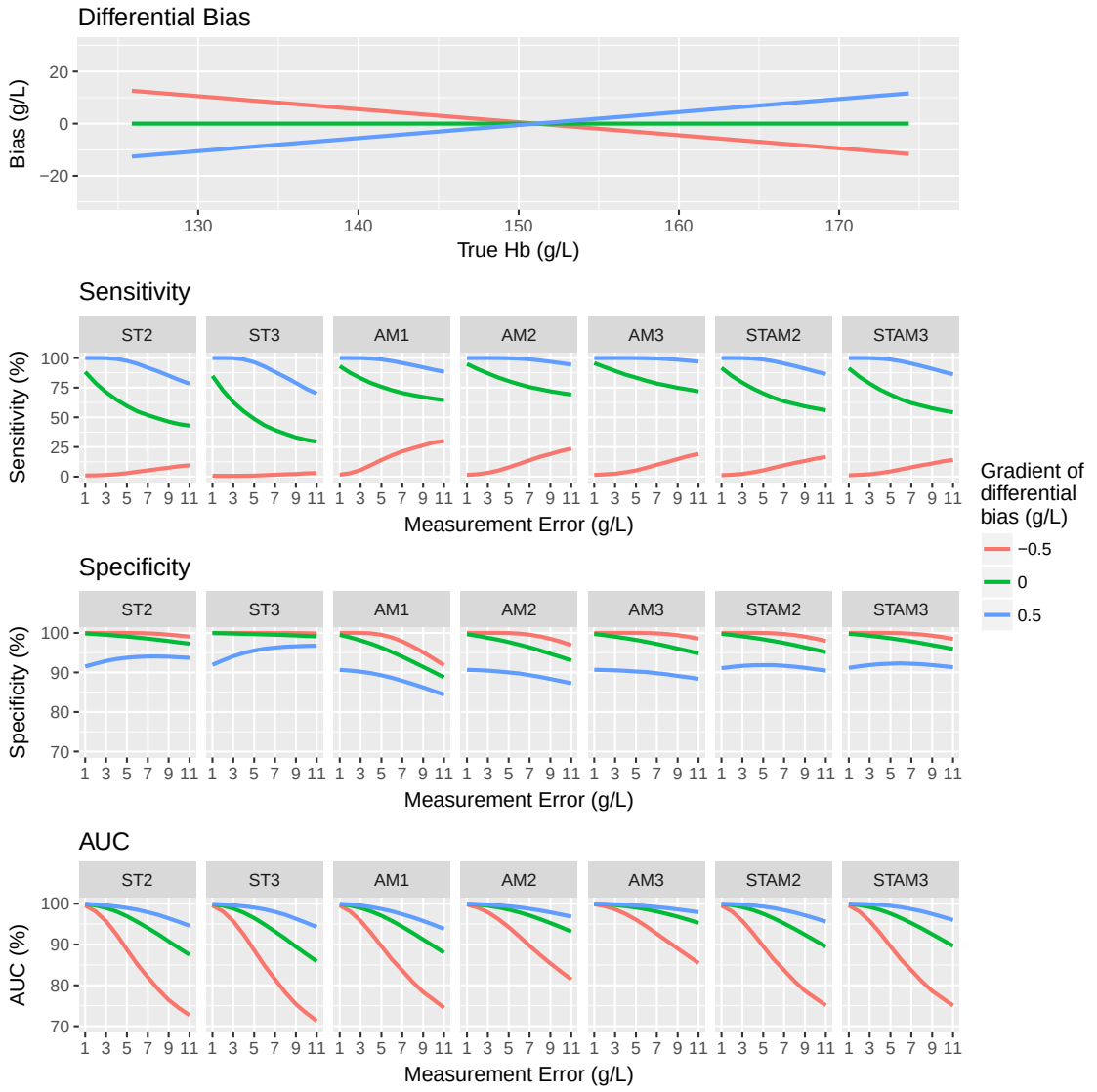

Figure 5 **Top**: Differential biases with varying gradients that were imposed for each testing strategy. The average device bias is 0 for all gradients. **Bottom 3**: The sensitivity, specificity and AUC exhibited for each testing strategy, according to the amount of differential bias and the amount of measurement error.

Strategy	Measurement Error (g/L)	AUC % (95% CI)	Coverage (%)	Mean Bias (g/L)	SE (g/L)		Average No. Measurements
					Mean	Empirical	
ST2	1	99.9 (99.8, 100.0)	95.0	0.03	0.20	0.20	1.06
	3	99.0 (98.6, 99.4)	90.3	0.13	0.21	0.21	1.06
	5	96.9 (96.0, 97.9)	65.5	0.34	0.22	0.22	1.08
	7	94.0 (92.4, 95.6)	13.8	0.71	0.23	0.23	1.10
	9	90.8 (88.5, 92.3)	0.1	1.22	0.25	0.25	1.12
	11	87.5 (84.7, 90.2)	0.0	1.86	0.27	0.27	1.14
ST3	1	99.9 (99.8, 100.0)	94.8	0.04	0.20	0.20	1.11
	3	98.8 (98.4, 99.3)	86.8	0.17	0.21	0.21	1.11
	5	96.5 (95.5, 97.5)	44.8	0.45	0.21	0.21	1.12
	7	93.2 (91.5, 94.9)	2.0	0.90	0.22	0.22	1.14
	9	89.5 (87.1, 91.8)	0.0	1.53	0.24	0.24	1.16
	11	85.9 (83.0, 88.7)	0.0	2.33	0.25	0.27	1.19
AM1	1	99.9 (99.8, 100.0)	94.6	0.02	0.20	0.20	1.00
	3	99.0 (98.6, 99.3)	94.4	0.02	0.21	0.21	1.00
	5	97.0 (96.2, 97.9)	94.6	0.02	0.22	0.23	1.00
	7	94.3 (92.9, 95.8)	94.8	0.02	0.24	0.25	1.00
	9	91.2 (89.2, 93.2)	94.6	0.02	0.27	0.27	1.00
	11	88.0 (85.5, 90.5)	94.3	0.02	0.30	0.30	1.00
AM2	1	99.9 (99.9, 100.0)	94.8	0.02	0.20	0.20	2.00
	3	99.5 (99.3, 99.7)	95.2	0.02	0.21	0.20	2.00
	5	98.5 (98.0, 99.0)	95.7	0.02	0.21	0.21	2.00
	7	97.1 (96.3, 98.0)	96.0	0.02	0.22	0.22	2.00
	9	95.3 (94.0, 96.5)	96.0	0.02	0.24	0.23	2.00
	11	93.2 (91.5, 94.8)	95.9	0.03	0.25	0.25	2.00
AM3	1	100.0 (99.9, 100.0)	94.8	0.02	0.20	0.20	3.00
	3	99.7 (99.5, 99.8)	94.8	0.02	0.20	0.20	3.00
	5	99.0 (98.7, 99.4)	94.9	0.02	0.21	0.21	3.00
	7	98.1 (97.5, 98.7)	95.3	0.03	0.22	0.21	3.00
	9	96.8 (95.9, 97.7)	95.9	0.03	0.23	0.22	3.00
	11	95.3 (94.0, 96.5)	95.7	0.03	0.24	0.23	3.00
STAM2	1	99.9 (99.9, 100.0)	94.8	0.03	0.20	0.20	1.06
	3	99.2 (98.8, 99.5)	93.5	0.07	0.21	0.21	1.06
	5	97.5 (96.7, 98.4)	86.2	0.18	0.22	0.22	1.08
	7	95.1 (93.7, 96.6)	66.0	0.36	0.23	0.24	1.10
	9	92.4 (90.3, 94.4)	30.8	0.62	0.25	0.25	1.12
	11	89.5 (86.9, 92.0)	7.6	0.94	0.27	0.27	1.14
STAM3	1	99.9 (99.9, 100.0)	94.8	0.03	0.20	0.20	1.11
	3	99.2 (98.8, 99.5)	92.9	0.09	0.21	0.21	1.11
	5	97.6 (96.7, 98.4)	83.6	0.22	0.22	0.22	1.12
	7	95.2 (93.8, 96.7)	55.0	0.43	0.23	0.23	1.14
	9	92.5 (90.5, 94.5)	17.7	0.72	0.25	0.25	1.16
	11	89.6 (87.1, 92.2)	1.6	1.10	0.27	0.27	1.19

Table 1 Comparison of AUC performance, coverage, mean bias, as well as the mean SE and empirical SE for the testing strategies. The average number of measurements for a population of donors using each testing strategy is also recorded, as the measurement error is varied between 1g/L and 11g/L

